Investigating Cultural Alignment of Large Language Models

Year: 2024 | Citations: 104 | Authors: Badr AlKhamissi, Muhammad N. ElNokrashy, Mai Alkhamissi

---

Abstract

The intricate relationship between language and culture has long been a subject of exploration within the realm of linguistic anthropology. Large Language Models (LLMs), promoted as repositories of collective human knowledge, raise a pivotal question: do these models genuinely encapsulate the diverse knowledge adopted by different cultures? Our study reveals that these models demonstrate greater cultural alignment along two dimensions -- firstly, when prompted with the dominant language of a specific culture, and secondly, when pretrained with a refined mixture of languages employed by that culture. We quantify cultural alignment by simulating sociological surveys, comparing model responses to those of actual survey participants as references. Specifically, we replicate a survey conducted in various regions of Egypt and the United States through prompting LLMs with different pretraining data mixtures in both Arabic and English with the personas of the real respondents and the survey questions. Further analysis reveals that misalignment becomes more pronounced for underrepresented personas and for culturally sensitive topics, such as those probing social values. Finally, we introduce Anthropological Prompting, a novel method leveraging anthropological reasoning to enhance cultural alignment. Our study emphasizes the necessity for a more balanced multilingual pretraining dataset to better represent the diversity of human experience and the plurality of different cultures with many implications on the topic of cross-lingual transfer.